ITEA Journal 2010; 31: 39–47

Development of Domain-Specific Scenarios for Training and Evaluation of Two-Way, Free Form, Spoken Language Translation Devices

Brian A. Weiss
National Institute of Standards and Technology, Gaithersburg, Maryland
Marnie Menzel
Appen Pty Limited, Chatswood, New South Wales, Australia

To create effective and accurate two-way, free form, spoken language translation devices, the technologies must have appropriate training data. The goal of the Defense Advanced Research Projects Agency Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is to demonstrate capabilities to rapidly develop and field this technology, so speakers of different languages can communicate in real-world tactical situations. A critical component is to generate data sets to train and evaluate the technologies. A novel approach was developed to collect these data, employing innovative data-collection and evaluation scenarios. This article describes the scenario methodology used for the TRANSTAC data collections and evaluations.

Key words: Communication in tactical situations; data collection; evaluation; Iraqi Arabic speakers; language translation systems; spoken language; U.S. military personnel.

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Projects Agency (DARPA) advanced technology research and development program. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field free-form, two-way translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter. To date, several prototype systems have been developed for specific domains in Iraqi Arabic (IA), Mandarin, Farsi, Pashto, and Thai.

The primary use cases involve U.S. military personnel and IA speakers, *Figure 1*. While the U.S. military personnel will be trained in advance to use the technology, the assumption is that the foreign language users will have little or no chance to become familiar with the system.

An Independent Evaluation Team (IET) was funded by DARPA to evaluate the TRANSTAC technologies during several phases of the TRANSTAC Program. The IET was responsible for analyzing the performance of the TRANSTAC systems by producing training data for the technologies, along with designing and executing technology evaluations and analyzing the results of the evaluation (Weiss et al. 2008).

This article discusses the initial approaches to collecting data and evaluating systems using the previous means of data-collection and evaluation scenarios. Further, subsequent novel approaches are presented for creating the audio, transcription, and translation training data through the development and execution of specialized data-collection scenarios. These innovative data-collection scenarios enabled the IET to collect more natural, tactical conversations that lasted 60% longer than their previous counterparts. In addition, the unique methodology for producing representative data and developing relevant evaluation scenarios is discussed. Comparisons demonstrate how the enhancements contributed to higher quality data and evaluation protocols. Ultimately, these scenario enhancements contributed to a technical

ITEA Best Paper from the Annual ITEA International Symposium, September 28–October 1, 2009, Baltimore, Maryland.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **MAR 2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Development of Domain-Specific Scenarios for Training and Evaluation of Two-Way, Free Form, Spoken Language Translation Devices**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology, 100 Bureau Drive,Gaithersburg,MD,20899**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

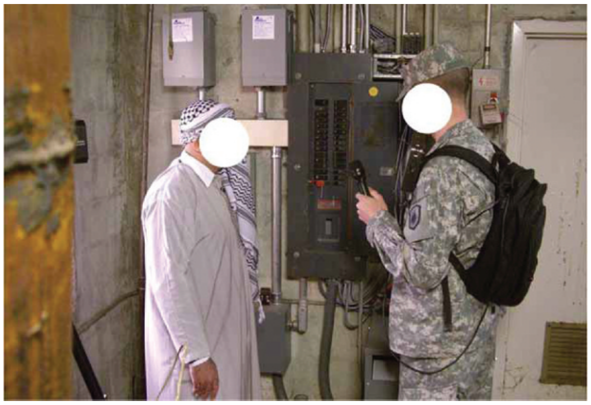


Figure 1. Demonstration interaction between a U.S. Soldier and an Iraqi Arabic speaker using a Translation System for Tactical Use (TRANSTAC) system.

performance improvement of 18% (discussed in detail later in this article).

Evaluation background

An experimental method was designed to evaluate the TRANSTAC technologies given their expected state of maturity. The IET developed an evaluation approach that would scale well with the technologies as they developed, thus allowing for valid assessments of performance improvements over time. This method included developing a scalable testing approach, securing participants for testing, etc. (Schlenoff et al. 2007).

The scalable testing approach was built off previous approaches and incorporated new procedures and evaluation types. The following two metrics were specified as the focus for the technology evaluation:

1. System usability testing—providing scores to the capabilities of the whole system,
2. Software component testing—evaluating individual components of a system to see how well they performed in isolation.

The IET employed a two-part test methodology to produce these metrics. The first metric was realized through the use of structured and utility evaluation scenarios, while the second metric was evaluated through the use of offline utterances. Both of these evaluation approaches were designed to measure the progressive development of the TRANSTAC system capabilities and to predict the impact these technologies will have on warfighter performance across a range of tactical domains. The scenarios' content was designed to provide a reasonable level of difficulty for the TRANSTAC systems at their current state of development as well to create the opportunity to evaluate the systems in the future at their expected rates of improvement. This methodology will be discussed further in subsequent sections. It should also be noted that individual technology performance scores are not published in this article due to restrictions in the DARPA TRANSTAC program.

Need for data-collection scenarios

In order to effectively assess the technology performance, conversational audio data between English-speaking military and IA personnel were recorded, transcribed, translated, and distributed to the technology developers prior to the evaluation, so that this data could be used to train the technologies. These conversations were driven by operationally relevant data-collection scenarios provided to each speaker at these collections (separate events held months prior to the evaluations).

These data-collection scenarios inspired the dialogues that allowed the teams to train their systems. Furthermore, a small portion of this audio data (known as the representative set) was not provided to the developers so that it could be used by the evaluation team to develop the evaluation scenarios. Utterances from the representative set were used to perform offline evaluations of the technologies that focused on the testing of automated speech recognition, machine translation, and text-to-speech. In addition, the representative set supported the creation of live evaluation scenarios that involved English and IA speakers interacting with the technologies.

Need for live evaluation scenarios

A specific evaluation scenario format and methodology was developed to evaluate both technical performance and utility of multiple technologies, so that comparisons could be drawn among them. These scenarios were generated directly from the representative set to ensure the systems were evaluated against relevant dialogues with which the developers were likely to train their systems.

Data-collection scenario development

Developing scenarios for the data collection included a series of steps, each necessary to ensure the creation of domain-specific and tactically relevant scenarios. A specific process was first adopted by the IET. As the program continued, the IET identified areas for improvement and enhanced the data-collection scenarios. Advantages of the new data-collection scenarios were identified at the conclusion of the data collection and evaluation events.

Initial design

The initial data-collection scenario design process began with taking inventory of the existing scenarios, followed by identifying new and pertinent topics. Next, the scenarios were generated following a specific format.

Inventory of existing scenarios. Since the goal of the TRANSTAC program is to develop translation capabilities that are relevant in real-world tactical situations, the starting point for the data-collection scenario design was to identify those scenarios, developed to support previous efforts, that were either highly applicable or underrepresented. Twenty existing scenarios were identified that either had not been recorded, had been used in a relatively small number of recordings, or were used in the previous phase's evaluations. These scenarios fell into the following topic areas:

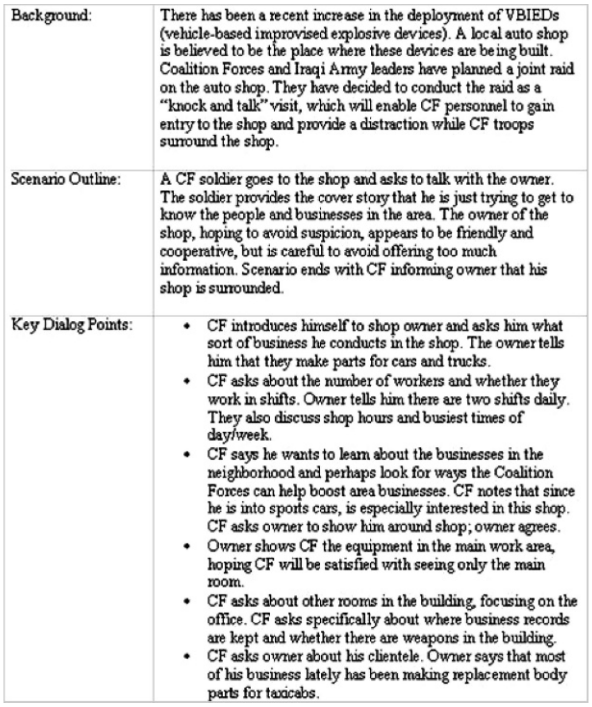
- Tactical Operations,
- Civilian Interactions,
- Joint Training/Operations,
- Intelligence Operations.

Topic identification. The first source of information for new topics originated from focus-group sessions involving military personnel of the target end-user population. IET personnel conducted discussions with Soldiers and Marines at numerous military bases around the country. These multi-day events began with conversations discussing experiences and incidents. IET members organized these topics, and dialogues were collected from this exercise to enhance the depth of each scenario topic.

The second source of information to further topic identification was to search media articles, testimonials, and relevant Web sites. This search enabled the IET to stay current on the interactions that occurred between English-speaking military personnel and IA speakers (both civilian and military).

Scenario creation. Once the information was organized, scenarios took shape through the development of one to two paragraphs that provided an overview and background along with five to 10 bullets that described the important concepts to be discussed in the dialogue. The entire scenario description was targeted to be approximately a single page.

In some cases, input from the end-user focus groups needed to be modified to make the material more suitable for data collection purposes. Specifically, topics that encouraged the English speaker to talk at length were avoided or reworked, so that reasonable opportunities were provided to the IA speaker to contribute to the conversation.



Background:	There has been a recent increase in the deployment of VBIEDs (vehicle-based improvised explosive devices). A local auto shop is believed to be the place where these devices are being built. Coalition Forces and Iraqi Army leaders have planned a joint raid on the auto shop. They have decided to conduct the raid as a "knock and talk" visit, which will enable CF personnel to gain entry to the shop and provide a distraction while CF troops surround the shop.
Scenario Outline:	A CF soldier goes to the shop and asks to talk with the owner. The soldier provides the cover story that he is just trying to get to know the people and businesses in the area. The owner of the shop, hoping to avoid suspicion, appears to be friendly and cooperative, but is careful to avoid offering too much information. Scenario ends with CF informing owner that his shop is surrounded.
Key Dialog Points:	• CF introduces himself to shop owner and asks him what sort of business he conducts in the shop. The owner tells him that they make parts for cars and trucks. • CF asks about the number of workers and whether they work in shifts. Owner tells him there are two shifts daily. They also discuss shop hours and busiest times of day/week. • CF says he wants to learn about the businesses in the neighborhood and perhaps look for ways the Coalition Forces can help boost area businesses. CF notes that since he is into sports cars, is especially interested in this shop. CF asks owner to show him around shop; owner agrees. • Owner shows CF the equipment in the main work area, hoping CF will be satisfied with seeing only the main room. • CF asks about other rooms in the building, focusing on the office. CF asks specifically about where business records are kept and whether there are weapons in the building. • CF asks owner about his clientele. Owner says that most of his business lately has been making replacement body parts for taxicabs.

Figure 2. Phase 2 data-collection scenario.

The data-collection scenario creation process led to the development of 44 scenarios that were deployed in varying frequencies across approximately six data collections. Before being used in the data collections, the scenarios were approved by a committee composed of military subject matter experts, IA cultural advisors, and evaluation team personnel. *Figure 2* shows a scenario generated to support the initial data collections.

Shortcomings. After these scenarios were put into practice, several deficiencies were noted. The first dealt with the organization of the scenarios within the four specified topic areas. Some topic areas contained many more scenarios compared with others. For example, only five of the 44 data-collection scenarios were categorized as Intelligence Operations. This led to an uneven distribution in scenario recordings and representation of dialogues across the four domains.

In addition, it was noticed that the data-collection scenario format imposed unintentional restrictions on the speakers. This included inhibiting a speaker's ability to enhance the dialogue with their own relevant experiences. The five to 10 bullets that were included in each scenario specifically laid out the flow of the scenario that prevented the speakers from augmenting the dialogue where they saw fit. Data collection participants also noted that recording scenarios from this format, especially after they were rehearsed, became somewhat monotonous and reduced the potential realism of the conversation.

Scenario evolution

Inventory of existing scenarios. Subsequent data-collection scenario development began the same way it did previously by taking inventory of the most recent scenarios. It was observed that the four scenario topic areas should be reorganized to better promote a more uniform distribution of scenarios to avoid the unevenness present earlier.

Topic identification. The evolution of this step began with more U.S. Marines and Soldiers being solicited for experiences to broaden the range of topics and to stay current with the types of interactions occurring among IA speakers. This information was collected from focus groups, discussions, and tactical training observations at military facilities including the Joint Readiness Training Center and the National Training Center.

The four topic areas were reorganized into six new domains to achieve a more balanced scenario distribution (among them) and to reflect the current interactions between English and IA speakers that would make the TRANSTAC technology more immediately useful.

A. Traffic Control Points/Vehicle Checkpoints,
B. Facilities Inspections,
C. Civil Affairs,
D. Medical Operations,
E. Combined Training,
F. Combined Operations.

With these new scenario domains in place, the groundwork is laid for the data-collection scenarios to evolve from their previous state.

Scenario creation. The data-collection scenario format evolved to encourage the speakers (both English and IA) to introduce more of their own experiences into the dialogues. Another intent was to provide a range of ideas and questions that the English speakers could choose in the event that they found difficulty in discussing a particular topic. In essence, the goal was to achieve a balance between providing the speakers with enough ideas to maintain a realistic dialogue (without them running out of things to say) and ensuring that the speakers would not get bored by reading near-scripted scenarios.

The new format began with organizing the ideas (along with the content from the previous scenarios that was still applicable) into the six scenario domains. For each domain, multiple English-speaker motivations were specified that are composed of background and situational information about the scenario. Following each English motivation, talking points were listed to give the speaker topics that could be included in their dialogues. Specific backgrounds and motivations were written for the IA speakers that fell within each of the English motivations. This process produced 60 viable scenario variants that were all used in the data collection recording sessions, *Table 1*.

Table 1. Phase 3 data-collection scenarios.

Domain	English motivation	Iraqi Arabic motivations
A – Traffic Control Point/Vehicle Checkpoint	1 – Quiet	4
	2 – IED hotspot	3
	3 – Border	3
B – Facilities inspection	1 – Police station	2
	2 – Power plant	2
	3 – Water treatment facility	2
	4 – Hospital	4
C – Civil Affairs	1 – Civilian complaint	3
	2 – SWET survey	3
	3 – Contractor interview	3
D – Medical	1 – MEDCAP/ DENCAP	3
	2 – Patient status	3
	3 – Medical attention	2
E – Joint Training	1 – Weapons	2
	2 – Patrols	2
	3 – Personnel/vehicle search	2
	4 – First aid	2
	5 – Patrol debrief	2
	6 – Arrest/detention	2
F – Joint Operations	1 – Planning a raid	4
	2 – Cordon and knock	4
	3 – Snap VCP planning	3
	Total scenario variants	60

IED, improvised explosive device; SWET, sewer, water, electricity, trash; MEDCAP, Medical Civic Action Program; DENCAP, Dental Civic Action Program; VCP, vehicle checkpoint.

Data collections

The data-collection scenarios were employed at the data collections. These events brought English-speaking military personnel and IA speakers together at a recording studio to generate audio dialogues. These two-way conversations were interpreter mediated since both the English and IA speakers spoke in their native languages (this also ensured that the dialogues would be smooth and succinct).

Prior to recording their dialogues, the speakers familiarized themselves with their data-collection scenario with the help of an IET member. From

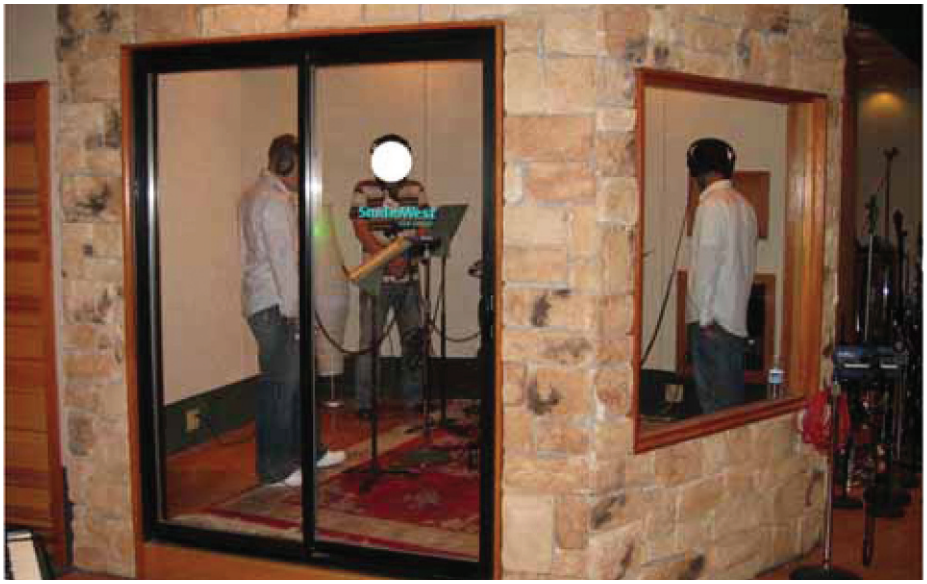


Figure 3. Speakers role-playing a scenario within a recording booth for data collection.

there, the three speakers (English, interpreter, IA) generated their dialogues inside a recording booth, *Figure 3*. Conversations varied in length from as short as 10 minutes to over 30 minutes. Conversation durations were largely at the discretion of the speakers based upon the scenario and their specific experiences.

Each weekend (consisting of two, 9- to 10-hour workdays) yielded from 18 to 36 hours of audio data, which were subsequently transcribed and translated. Producing large volumes of data was accomplished by having up to three recording sessions run in parallel.

Advantages of design improvements

The innovations in the data-collection scenarios encouraged the speakers to be more engaged, which ultimately resulted in more comprehensive and longer-lasting dialogues. The speakers commented that the format enabled them to inject more of their own experiences and provided them reasonable latitude to take the dialogue in a familiar direction, while still staying within the assigned motivation.

Since the innovative format allowed the speakers to become more immersed in their dialogues, the IET collected an average of 8 minutes more per scenario. This is based upon an average of 20.70 minutes per scenario (580 scenarios across 200.1 hours of recordings) compared with 12.47 minutes per scenario (637 scenarios across 132.4 hours of recordings) collected during previous recordings. This additional data not only benefited the technology developers, since they have more data to train their systems, but also provided the IET with a richer and larger data set to support the evaluations.

Live evaluation scenario development

The development of the live evaluation scenarios took shape after the data-collection scenarios were transcribed and translated. This process began with splitting the data into two pieces (per weekend collection event), where a majority went to the developers for training and the remaining portion stayed with the IET for the evaluations. Next, specific scenarios were selected and adapted to be used in the evaluations.

Initial design

Representative data set development. Before the involvement of the National Institute of Standards and Technology (NIST), the data that had been withheld for evaluation purposes were selected solely on the basis of scenario and demographic information. The particular features given priority in this selection process were as follows (in order of importance):

1. Scenario Type,
2. Speaker Dialect,
3. Speaker Gender,
4. Speaker Age.

The main drawback to this approach was that the withheld data only represented the total data set at a scenario/demographic level and not at a word level. In order to address this concern and produce a set of scenarios that were more representative of the training data, the following approach was proposed:

1. Collect all instances of all words in the training set (available at the time).

2. Take out very common words (e.g., "the," "of," "I," in English).
3. For a given scenario, determine
   a. the percentage of words that are in scenario that are also in training set, and
   b. the average number of times a word in the scenario appears in the training set.
4. Select the scenarios with the highest combined average score.

This approach was applied using the initial set of data, which consisted of 118 sessions (approximately 29 hours of data). Statistical word analysis[1] was performed on this data set according to the following categories:

a. *Total words in scenario,*
b. *Total unique words in scenario,*
c. *Words only occurring in this scenario,*
d. *Unique words only occurring in this scenario,*
e. *Percentage of words common to this scenario and other scenarios,*[2]
f. *Percentage of unique words common to this scenario and other scenarios,*[3]
g. *Average number of times a word in the scenario appears in the training set.*

Approximately 10% of the total set of scenarios selected as being the most representative consisted of those with the highest combined average, based on the sum of *(f)* and *(g)*.

The resulting set showed high coverage of the words in the training set but did not provide a representative distribution of scenario topics, that is, some scenarios were significantly overrepresented, while others were not represented at all. To rectify this, some of the repeated scenarios were replaced with alternative scenarios that shared a similarly high score.

It was then proposed that scenarios with mid-range scores may be more representative than those with higher scores as they would facilitate maximum diversity in scenario types, but this would come at the cost of an increase in the number of words appearing in the evaluation data but not in the training data.

It became evident that the percentage of unique words may have a more significant bearing on achieving comprehensive representation than the number of times a word shows up in the training set. As a result of this finding, a different approach (based on unique word metrics) was developed to select the representative set. The procedure was as follows:

1. Sort the scenarios by *Percentage of unique words common to this scenario and other scenarios.*
2. Take the middle 30% (or so) of the scenarios.
3. Take this new set of scenarios and sort first by *scenario number* and then by *average number of times a word appears in the training set.*
4. The next step is somewhat subjective. Take at least one instance of each unique scenario number while trying to include a wide-ranging distribution of English speakers and maximizing the number of times that words in the scenario appear in the training set.

This process was applied to the subsequent batches of data and successfully produced representative sets in each instance.

Scenario selection. For the first evaluation, the technologies were tested against 9 live field[4] and 11 live lab[5] scenarios. First, scenarios were selected from the representative set, which were adapted to be field-structured scenarios (discussed in more detail in the following section). This decision was made first because the field evaluation presented a more constrained environment as compared with the lab evaluations with regard to scenario realization. The initial step in choosing the field scenarios was to sift through the representative set to determine which scenarios could be realized in the limited field environment (Weiss et al. 2008). The next step was to pick nine scenarios from the "field acceptable" set, emphasizing a representative balance of scenarios from each of the four topic areas.

Next was to choose the scenarios for use in the lab evaluations. Since the field evaluations included solely structured scenarios, while those from the lab consisted of both structured and scripted,[6] the lab-structured (structured scenarios that occurred in the lab environment) were determined first to avoid repeating the field-structured. evaluations. Based upon these criteria and the unselected scenarios remaining in the representative set, six scenarios were selected to be lab-structured scenarios.

Selecting the scenarios to be scripted for the lab was done by scrutinizing the representative set with a focus on those scenarios that had yet to be chosen for the evaluations and those with "clean" dialogues. It was desired to make the lab-structured and scripted scenarios unique from one another while it was acceptable to use the same scenario in the field and the lab-scripted (scripted scenarios that occurred in the lab environment). This could be accomplished since all field scenarios were structured, making them unique in evaluation dialogue. This resulted in five scenarios being scripted for the first evaluation.

The second evaluation also contained 20 live evaluation scenarios, but it consisted solely of structured scenarios (the scripted scenarios were removed from the evaluation since they were unnatural for the speakers and turned out to be minimally repeatable).

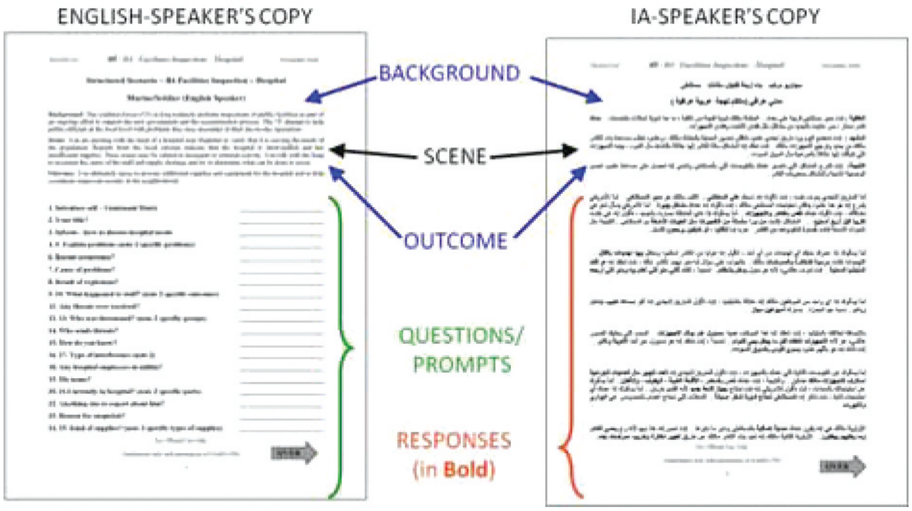


Figure 4. Structured evaluation scenario showing both speakers' sheets.

This evaluation included several scenarios that were used in the first test event to see how far the technologies had improved.

Dialogues were selected to be field- and lab-structured scenarios in a manner similar to what was done in the first evaluation. For those selected scenarios that were used in the first evaluation, their past write-ups were modified and augmented with additional English prompts and IA responses based upon the current representative data.

Scenario adaptation. Adapted from the representative set, structured scenarios were intended to prompt the English-speaker to ask the IA speaker questions to determine information known to the IA speaker. Both the English and IA speaker's scenarios outlined the scenario background, set the scene, and presented the scenario's intended outcome. The English speaker's scenario continued with numbered prompts directing them as to the specific pieces of information they were to gather or knowledge they were to pass to the IA speaker. Instead of prompts to specific pieces of information, the IA speakers were provided with several paragraphs outlining the information they were supposed to convey when appropriately queried. *Figure 4* presents an evaluation scenario.

The scripted scenarios were very straightforward in design. Of the remaining scenarios, the five with the "cleanest" utterances were chosen. From 24 to 29 utterances (both English and IA) were selected per scenario to form the five scripted scenarios. If a selected utterance contained a mispronunciation (e.g., "ah," "um") then the utterance was cleaned up (the mispronunciation was eliminated) for the scripted document. However, any slang or street terms were retained.

Evaluations using live scenarios. The technology teams were tested against each of the live structured evaluation scenarios in 10-minute time frames (per scenario). For a structured scenario, the IET measured how many concepts the English speaker obtained from the IA speaker within 10 minutes. Likewise, the IET used low level concept transfer metrics along with Likert and automated metrics to evaluate the scripted dialogues (Sanders et al. 2008).

Shortcomings. In addition to the scripted scenarios being removed from the test plan between the two evaluations, the other significant shortcoming of the structured scenarios involved their use within the field environment. The English and IA speakers noted that they felt more immersed in the scenarios within the field as opposed to the lab (the high-level concept transfer metrics supported a greater speaker comfort in the field, as well). However, the speakers felt very constrained using the structured scenarios within the more realistic field environments. This information was gathered after debriefing the speakers at the conclusion of the evaluations. The IET noted that it was still important to continue the lab evaluations (for comparison to previous evaluations to see how the technologies had advanced with respect to high-level concept transfer) but realized that the scenario format needed to be altered for the field tests.

Subsequent evolution

The following evaluations not only continued the structured scenario format within the lab environments but also introduced scenarios specifically designed for the field to assess the end-users' utility of the technology. These so-called utility-field scenarios were

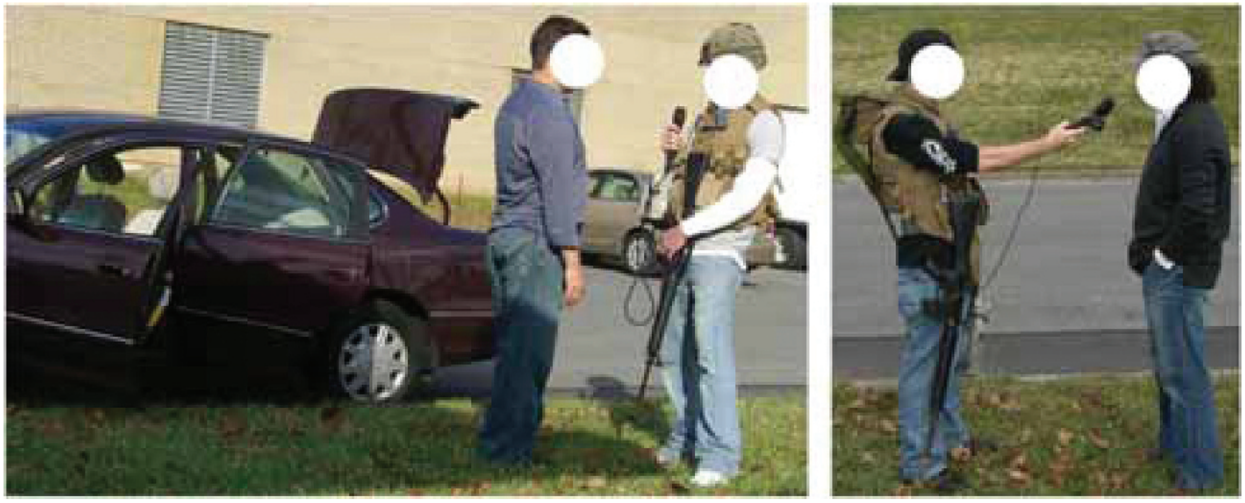


Figure 5. Phase 3 field evaluations with dialogues motivated by the utility scenarios.

less constrained and took a format similar to the data-collection scenarios. Three data-collection scenarios (that were also selected to be lab-structured scenarios) were chosen to be performed as utility scenarios in the field. The scenario selection was based upon which domains and corresponding English and IA speaker motivations were realizable in the available field setting. Once these motivations were chosen, the English speaker's talking points were adjusted to better reflect the test environment.

The final product yielded scenarios in which each English speaker was provided with a motivation and talking points (similar to that provided to the speakers during the data-collection events). Likewise, each IA speaker was provided with their own background and motivation. *Figure 5* depicts some English and IA speakers interacting in more realistic environments as prescribed by the utility scenarios.

Evaluation format progression benefits

Direct technical performance comparisons were drawn across the technologies over multiple evaluations when tested against the structured scenario format. Likewise, the utility format enabled the speakers to use the system in somewhat realistic/tactical manners where they assessed the utility of the technologies.

In addition, the IET noted an approximately 18% (on average) increase in the high-level concept transfer metric between the second and third evaluations (that NIST conducted) across three principal TRANSTAC technologies. Although numerous factors affected the teams' improvement across the phases, including their accessing more training data and having more time to enhance their technologies, the evolution of the evaluation scenarios played an important role. The structured scenario and utility field scenario formats enabled the IET to specifically assess technical performance and utility of the TRANSTAC systems, respectively, while allowing direct comparisons among the systems to be established.

Future efforts

The program is continuing to move forward, and the evaluation team is further refining the design and implementation of data-collection and evaluation scenarios to support both technical performance and utility tests.

Conclusion

The data-collection scenario development process enabled the evaluation team to collect tactically relevant, realistic dialogues between English-speaking military personnel and Iraqi Arabic speakers. As a direct result of this effort, the research teams were provided with appropriate data with which to train their systems, and the evaluation team was given representative audio with which to generate fair and appropriate evaluations. The evaluation scenario development process enabled the IET to create appropriate scenarios such that multiple technologies were evaluated and compared following conversations between English and Iraqi Arabic speakers, while adhering to structured and utility-field scenario formats.

It is important to note that these processes of creating and implementing both data-collection and evaluation scenarios can be applied to the training and evaluation of spoken language translation devices focused on languages other than English and Iraqi

Arabic. In addition, scenarios can be designed that are outside of tactical military dialogues.[7] ❑

Brian Adam Weiss has been a mechanical engineer at the National Institutes of Standards and Technology in Maryland since 2002. His focus is the development and implementation of performance metrics to quantify technical performance and assess end-user utility of intelligent systems throughout various stages of development. His current projects include assessments of soldier-worn sensor systems and spoken language translation devices. He has a bachelor of science in mechanical engineering from the University of Maryland, a professional master of engineering degree from the University of Maryland, and is working towards his doctor of philosophy in mechanical engineering with the University of Maryland. E-mail: brian.weiss@nist.gov

Marnie Menzel is a project manager/linguist at Appen, a Sydney-based provider of speech and language technology resources. Marnie joined Appen in 2005 and has led a number of very large scale multi-national projects for both commercial and government organizations. Marnie holds a bachelor of arts degree (Hons) in linguistics from the University of New South Wales in Sydney, Australia. Her research interests are in human interaction in a conversation analysis framework focusing on question/answer sequences. In the context of the DARPA TRANSTAC program, Marnie has applied her research to the development of interactive approaches to speech data collection, which optimize naturalness and authenticity. E-mail: mmenzel@appen.com.au

Endnotes

[1] Due to time constraints, common words were not removed when generating the statistics.

[2] If the word "checkpoint" is used 20 times in a scenario out of the total 1,000 words in a scenario, and "checkpoint" is present in other scenarios (whether it's one more or many more), then this would account 20/1,000 = 2% of the *words common to this scenario and others*.

[3] If the word "checkpoint" is used 20 times in a scenario out of the total 1,000 words with 250 unique words total in the scenario, and "checkpoint" is present in other scenarios (whether it's one more or many more), then this would account for 1/500 = 0.4% of the *unique words common to this scenario and others*.

[4] Live field evaluations were set up to test the systems in a more realistic environment. This included introducing very well-controlled background noise, requiring the English-speakers to carry the technology, and encouraging the speakers to be mobile during the evaluation.

[5] Lab evaluations were designed to test the systems in an idealistic environment, with no background noise and the participants being stationary.

[6] Scripted scenarios were dialogues taken directly from numerous data collection recordings that the speakers read verbatim into the technologies during the evaluation.

[7] NIST disclaimer. Certain commercial products and software are identified in this article in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

References

Sanders, Gregory, Sebastian Bronsart, Sherri Condon, and Craig Schlenoff. 2008. Odds of successful transfer of low-level concepts: A key metric for directional speech-to-speech machine translation in DARPA's TRANSTAC program. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference,* May 28–30 2008, Marrakech, Morocco. Paris, France: European Language Resources Association.

Schlenoff, Craig, Michelle P. Steves, Brian A. Weiss, Mike Shneier, and Ann Virts. 2007. Applying SCORE to field-based performance evaluations of soldier worn sensor technologies. *Journal of Field Robotics* 24 (8–9): 671–698.

Weiss, Brian A., Craig Schlenoff, Michelle P. Steves, Sherri Condon, Jon Phillips, and Dan Parvaz. 2008. Performance evaluation of speech translation systems. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference,* May 28–30 2008, Marrakech, Morocco. Paris, France: European Language Resources Association.

Acknowledgments

This work is funded through the DARPA TRANSTAC program, and the authors acknowledge and greatly appreciate the support of the TRANSTAC program manager, Dr. Mari Maeda.